Basbalo, Automized Baseball Article Generator: Based on Sabermetrics*

Seungwoo Schin and **Martin Ziegler**
principia_12@kaist.ac.kr

Abstract

Recently, robot journalism have been a big issue in natural language processing. However, most of the robot journalism frameworks appliced the static template method, which lacks variety of the text. In order to enhance diversity of generated text, this paper applies various natural language generation methods. As a result, this paper proposes framework for generating article that explains a baseball game. Overall process of framework suggested in this paper is as following. First, analyze game record to extract events in the match, and apply support vector machine to select events that should be included in the article. Second, generate paragraph structures based on the static template. Third, generate sentence using combinatory categorical grammar(CCG) based on modal logic. The resulting articles are evaluated by professioanl baseball journalists.

$\int exp(-\alpha x^2)$

*This document has been adapted from the instructions for earlier ACL and NAACL proceedings, including those for NAACL HLT15 by Matt Post and Adam Lopez, NAACL HLT12 by Nizar Habash and William Schuler, NAACL HLT10 by Claudia Leacock and Richard Wicentowski, NAACL HLT09 by Joakim Nivre and Noah Smith, for ACL05 by Hwee Tou Ng and Kemal Oflazer, for ACL02 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats. Those versions were written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

1 Introduction

Recently, the robot journalism is applied to various field, such as earthquake alarm (Times, 2014), stock market report generation, and sport match article generation (Allen et al., 2010). However, the robot journalism is not a new concept; according to the research by Asmodt & Nygard, 1995, correlation of data, information, and knowledge lead to advance of artificial intelligence managable integrated systems (Aamodt and Nygrd, 1995). Moreover, research by Bakker, 2012, predicted that low-cost, auto-generated content, including news articles, will be developed (Bakker, 2012).

Current advance of robot journalism is based on rapid advance of naturaul language processing, machine learning and large scale datasets. Natural language processing technology is responsible for *how* to transfrom given information into natural language. Machine learning, on the other hand, is responsible for *what* information should be contained in the text. Large scale dataset provides a foundation for machine learning and natural language processing.

Among various fields in natural language processing, current robot journalism is rooted on natural language generation(NLG) (Dorr, 2015). To be more specific, current robot journalism frameworks are mostly based on the static template method. In the static template method, natural language is composed by inserting context sensitive words to the static template. One application of static template method is the work of Allen, Templon, McNally, Birnbaum & Hammond, 2010, *StatsMonkey*, framework for baseball article generation (Allen et

al., 2010). The work of Kim & Lee, 2015, also presents the static template based approach, and applied NLG technology to korea baseball organization league data (Dongwhan Kim, 2015).

However, the static template method has certain limiations such as reflecting personal preference or variety in phrases (Dongwhan Kim, 2015). Limitations of the static template method lead to a need for robot journalism frameworks based on different NLG methods. Work of Hisar, 2003, is an exmaple for applying evolutionary algorithm to poetry generation. In this paper, NLG based on combinatory categorical grammar is applied.

To be more specific, this paper will propose a new framework consisting of three steps. First, raw data formalization according to internal database of framework. Second, event evaluation using support vector machine. Third, application of combinatoric categorical grammar and sentence structure extraction for generating sentence template. Our research used korea baseball organization league data to generate articles for explaining baseball games. Korean language is used in the dataset and resulting articles.

Remaining parts are organized as following. Section 2 covers summary of related work. In Section 3, theoratical background of openCCG and example sentence realizations are presented. In Section 4, framework structure from raw data to result article is presented. Seciton 5 contains quantitative evaluation on result articles. Section 6 concludes this paper by presenting summary and possible future work.

2 Methodology

2.1 Background

2.1.1 Natural Language Generation

Natural language generation(NLG) is a field of natural language processing devoted for generating natural languages in various forms. Usually generation process consists of two plans: tactical(deciding how to say it) and strategical(deciding what to say). The dual between strategy and tactic is realized as sentence generation and text planning. The structure of text in most NLG framework is thus a tree structure, which has leaves as each sentences and paragraph, text, or higher hierarchy containing lower hierarchy contents. Generation of actual text from abstract tree structure is called surface realization. Depending on how sentence planning, sentence generation, and surface realization is implemented, NLG techiniques can be classfied into four main categories(D.D.McDonald, 1992): canned text method, static template method, phrase-based method, feature-based method.

The canned text method is the simplest method for sentence generation. Setences are manually generated and hardwired in the framework. Despite the strength of simple implementation, canned method lacks flexibility. The static template method uses a template, which is set of sentences that have missing words. Resulting natural language is generated by inserting appropriate words in the template, and framework is responsible for what template to use and what words to use. Choice of template and appropriate words is a naive form of text planning and sentence generation, respectively.

The phrase-based method can be seen as a generalization of the static template method. Phrases structures are learned from training data or manually generated, and saved in the framework. Saved phrase structures are assembled to form a sentence based on the syntax of the main language. Therefore, in the phrase-based method, syntax of the target language need to be implemented in the framework.

The feature-based method models sentences as an unique set of features, probably organized in an abstract structe such as undirected graph. The feature-based method begins with collecting features from the input raw data. Then the framework makes a tree-structure plan for whole text from top to bottom. The plan is made by distributing features to be contained to each component of the text. Distribued features are assembled to make the resulting text from bottom up, which corresponds to surface realization.

In most practical frameworks, hybrid approaches are made. Currently, there are variety of NLG frameworks with different implementations and purposes. Representative NLG frameworks are (gen, 2012), (kpm, 2012), (cli, 2012), (nat, 2012), and (ope, 2012). In this paper, openCCG will be used as framework.

2.1.2 Combinatory Category Grammar

Combinatory category grammar(CCG) is one implementation of the feature-based method. In CCG approach, categorial grammar is used as a internal sturcture for modeling natural language. In formal language theory, grammar is a mapping from word chunks to abstract structures. Categorial grammar(CG) is one of such mappings; it maps words to categories.

The Category is an abstract structure. There are two kinds of category: atomic categories and complex categories. Atomic category is a set of simple, non-composite, atomic symbols that builds a foundation for CG. This is simliar to that of symbols in context free grammar. Part-of-sentence can be a good example of atomic category. Complex categories, on the other hand, is build from composing atomic categories. Build process is done by applying operators on the atomic category. In this paper, two operators are used: forward application operator and backward application operator. CCG is an extension of CG for more flexible surface realzation.

The surface realization of CCG is based on hybrid logic dependency semantics. CCG models sentences as directed graph, with nodes correspond to words and edges correspond to dependency between words. Words and dependency is labeled to graph. Dependencies in graph might be syntatic dependencey, general depndency, or domain-specific dependency. Although many sentences can be modeled as tree structure, there are no topological contraints to model graphs. A graph might be unconnected, contains loops, or non-tree.

The model of sentences can have different depth in abstrction and dependency to domain of generation. When model only contains syntactic dependency structure, the model is said to be surface sentence plan. Undirected graph can be realized directly from pre-defined grammar. On the other hand, when model contains general dependency, it is said to be deep sentence plans and need extended grammar or ontology to make model into a natural language. Tree model is realized to sentences using the formulation to logic formula. In openCCG, hybrid multimodal logic is used to enhance flexibility of sentence generation.

2.1.3 Sabermetrics

In baseball, records and statistics are treated extremely critical in analyzing each game. In baseball statistics, there are mainly two types of statistics: first order and second order statistics. First order statistics are raw data from baseball game, such as hit rate, number of home runs, etc. Second order statistics are built from first order statistics usually by applying artihematic operations on various first order statistics. Sabermetrics is the statistical approach for generationg second order statistics (James, 1985). After Bill James have introduced basic sabermetrics, it have been extremely successful in evaluating and predicting the behavior of baseball players. In this paper, sabermetrics provieds a foundation of modeling baseball games.

2.2 Implementation

In the following sections, implementation of proposed framework is explained in detail, from raw data to resulting text.

2.2.1 Raw Data Formalization

Baseball game in this framework is modeled as a sequence of states. States are basically snapshot of the baseball game at specific time. Since baseball games progress in unit of pitches, the states are expected to contain information of the current state of the game after every pitch. Transition between states are called events. Events contain information about what happened between state transition, and change first order statistics of game accordingly. For example, when a player hits and advances to first base, the hit event changes the state accordingly so that changed state reflectes the advnace of the hitter.

NLG frameworks begin with formalizing raw data inputs as desired format. Framework proposed in this paper accepts raw data from two sources: from text relay and manual record from official scorer. A Text relay for baseball game is provided by portal sites and Korea Baseball Organization(KBO). Even though this source lacks consistency and accuracy, the source is provided for every game from 2008. The other source is from official scorer from KBO, which is very accurate and informative. However, only available reocrds are from 2016 season. In order to build complete database, framework uses both data source to build internal database. Since two raw

data sources are different in format and information content, parser for each dataset is implemented independently.

Both parsers for each data source parses sequence of state and events from the raw data source. Extracted states and events provide ingridient for baseball article generation. However, not all events are used as ingridients due to the diffenrence of impact of each event. Such impact difference leads to necessity of evaluating each event.

2.2.2 Event and Game Evaluation

Events can be evaluated in several perspectives. In this research, mainly two types of event evaluation is made: impact on winning/losing and attention of the public. Impact on winning and losing can be determined by statistical analysis, and thus it is totally dependent on the game content. Public interest, however, does not only depend on the impact of the game. Some events does not effect game result while drawing big attention from audience.

Event evaluation based on impact to the game is conducted using support vector machine(SVM). Each event is vectorized using features provided by both sabermetrics and additional features. To be more specific, sabermetrics such as winning probability added(WPA), leverage index(LI), weighted runs created plus(wRC/wRC+) change are selected as feature. For the additional features, consideration on statistical history of each player is manually implemented to specify the exceptional events. For example, a steal of players with slow running speed can be treated as exceptional event. The SVM classifier classfies the event into two categories: suitable for article and not suitable for article.

The training data of the event for SVM is generated from event extraction from prior articles. Prior articles are crawled and tokenized into sentences. Tokenized sentences are than parsed to abstract syntax tree structure where the syntax have hardwired baseball-related part-of-speech in it. For example, name of player is not tagged as proper noun; instead, it is tagged as player part-of-speech. By constructing a syntax tree for toknized sentence, it is possible to extract the list of events that are contained in articles written by human journalists. Syntax trees constructed here will be later again applied to build a sentence generator.

The other part of evaluation is based on predicted attention of public audience. Since there is no statistical method to infer such event, case study about popular issues formed on KBO leauge games is conducted. From the case study conducted on records from 2008 to 2016, various types of exceptional evaluation criteria is made and classifiers for each criteria are implemented to detect such events.

Game type also needs to be specified, since the type of game contributes to the introduction section and overall mood of article. Game type is a vector containing features for various criterias, such as hitting-focused/pitching-focused game, win/lose state, rival match, ranking deciding match, etc. Each features are assigned with a boolean part and an impact part. The boolean part is responsible for whether the flag is on, and an impact part is responsible for evaluating the significance of the feature on deciding game type. When framework needs only one feature for the game type, only the feature with most impact will be considered and other features will be ignored.

2.2.3 Article Planning

With the selected events, plan for article structure can be constructed. In this framework, the structure of article is composed of four parts: title, introduction, content, and conclusion.

Introduction and Title Introduction part contains information about the game itself, such as date of the game, brief summary of the game, result of the game, first pitcher of each team, etc. Also, extremely exceptional events or performances of a player are also described here. Title is generated based on event or player with the most impact. This part of article is based on the static template type. The proposed framework contains static templates for possible types of game which can be generated by above game type decision process.

Content Content is the part for describing the selected events from above evaluation process. The content part is composed of several paragraphs that are timaly ordered. Each paragrpaph explains a turn or few turns. The paragraph contains sentences that explains few selected events, and summary sentence for the turn explained in the paragraph. Sentences containing information about selected evetnts are

generated using CCG and learned phrases. Details of sentence generation process is proposed in the following section.

Summary sentence will be generated to reflect the information of the explaining turn. The information of the turn is abstracted using the mood interface. The mood of a specific turn is calculated from the whole events in the turn. For examplem, if there was a reversal in the turn, the mood will reflect the reversal. In this framework, there are sixteen moods available, depending on the winning probability change of each teams. Mood acts as interface between actual events of the overall turn and summary sentence for the turn. Sentence generation of summary sentence is done by both using the static template method and phrase learning method.

Conclusion Conclusion part contains information about evaluation on performance of each team and key players. Also, it contains information about scheduled game and ranking change of each team. The framework also contains static template for conclusion for possible game types.

2.2.4 Sentence Generation

Sentence generation is done by hybrid method of mixing the static template method, phrase-based method and the feature-based method. Static templates are manually generated based on prior baseball game article and words are filled with appropriate selected words. Words are selected accordingly from framework's internal database.

The sentences are extracted from previous articles using part-of-sentence tagging and tag analysis. Each sentence is transformed into various syntax trees depending on the dependency of the baseball domain. By applying the baseball-dependendt part-of-speech and grammar, it is possible to extract game-free tree structure from sentences. Generated tree structures are than saved in the internal database labeled with its semantics. The semantics are implemented in hybrid multimodal logic formula.

Static templates generated in the framework have two kinds of templates: templates that acutally contains manually coded sentences in natural language and template that contains only the type of the sentence and its game-dependent content. Thus the static template can be genearlized to have flexible results. Such realization of static templates are used to generate introduction and conclusion section.

Generating sentence for content requires different approach. Since it is impossible to predefine the format of the sentences, sentence planning within the paragraph should be made. Sentence planning is basically grouping events that are allocated to the paragraph, so that paragraph can contain sentences that explains a given list of events. Sentence planning is done by considering mainly two features: the length of generated sentence and mood of the whole paragraph. After the grouping of selected evnets are finished, the framework assigns most suitable sentence structure. Abstract sentence structure is than realized to actual sentence written in natural language.

3 Results

In the results section, actual implementation result and specification of framework will be given. Also, discussion about the output of our framework will be covered. The framework in this paper generated articles explaining about 288 games in Korea Baseball Organization (KBO) league in 2015, 2016 season. For each game, the framework generated up to twenty articles depending on possible issues or bias that can be applied in the article. Whole framework is implemented using python3, mongodb, and openCCG.

3.1 Framework Implementation Result and Discussion

Implementation of the framework consists of 3 parts: corpus generation, article planning, and sentence generation. Performances of intermediate result made in each step are evaluated accordingly.

Corpus Generation In this study, the proposed framework contains five corpora: baseball article, baseball term, general Korean language corpus, syntax corpus, and template corpus. Baseball article corpus is generated by crawling data from various media. By crawling baseball related articles from web, corpus is organized and tagged with metadata: title, source media, writer, published date, and title. In this study, 121,387 articles are collected. All collected articles are tokenized into sentences and paragraphs.

Baseball term corpus contains various sets of words about KBO league. Terminologies related to baseball game rules are manually implemented using XML. Ontology of baseball terms are extracted from KBO official rule and terms (kbo, 2016). API provided from KBO record managing organization is used to generate corpus for terms related to teams and players: player names, team names, and stadium names are included. Gerenrated corpus contains 698 rule-related terms, 616 players, 10 teams, 21 stadiums, and synonyms of each item for enhancing expressive power of the framework.

For handling general Korean language, corpus provided by Korean Natural Language Processing with Python (KoNLPy) is applied. General language corpus is responsible for the words that above two corpora cannot handle.

The syntax corpus is also based on KoNLPy. However, it uses different part-of-speech (POS) ontology that is specialized for baseball related texts. For instance, name of player have POS called Player instead of proper noun. New POS ontology is generated based on baseball term corpus by overloading appropriate POS of each terms in baseball term corpus.

The template corpus contains static templates of sentences, paragraphs, and articles. Static templates in this corpus are manually generated and tagged in order to provide baseline performance of the framework. Framework consults static templates when sentence generation cannot be handled in other parts of the framework. The corpus contains sentence structure for possible types of an event and paragraph structure for possible types of a game.

Article Planning Evaluation In this framework, article plans are tree-structured schema for generating articles. In order to make an appropriate article plan for given match, database of previous article plans are generated. In this step of the framework, each paragraphs are converted to a tuple of a mood and a sequence of events contained in the paragraph. The accuracy of a generated database is manually checked by selecting random 100 articles, which showed 87% accuracy. (need to be replaced to table)

Sentence Generation Evaluation As in the article planning implementation, sentences from previous article corpus are parsed by generating syntax tree using the syntax corpus. Sentence structures for 3105 sentences from 100 articles are manually evaluated. (need to be replaced to table)

3.2 Framework Output Result and Discussion

The framework generates multiple aritlcles for each game, depending on the possible bias of the game. The evaluation of the result article is done by conducting survey to public readers. Participants of the survey are given articles about ten games. For each game, ten to twelve articles are listed, and user evaluates two articles in three criteria: Accuracy of the information, adequacy of the information, and expression in the article.

Accuracy of the Selected Information Tables will be added

Adequacy of the Selected Information Tables will be added

Expressive Power of Articles Tables will be added

4 Discussion and Conclusion

Result surveys pointed out that the generated language from Basbalo is rich in linguistic variety. Despite minor errors in written articles, resulting articles satisfies criteria for publication. As seen in the table, Basbalo showed outstanding performance in information selection, both in adequateness and accuracy. Also, the results show that resulting articles successfully mimic the article written by human journalists.

Acknowledgments

Thanks to professor woosuk Park in the KAIST, school of humanities and social sciences.

References

Agnar Aamodt and Mads Nygrd. 1995. Different roles and mutual dependencies of data, information, and knowledge an ai perspective on their integration. *Data and Knowledge Engineering*, 16(3):191 – 222.

Nicholas Allen, John Templon, Patrick McNally, Larry Birnbaum, and Kristian Hammond. 2010. Statsmonkey: A data-driven sports narrative writer.

Piet Bakker. 2012. Aggregation, content farms and huffinization. *Journalism Practice*, 6(5-6):627–637.

2012. Clint. http://www.cs.bgu.ac.il/ elhadad/clint.html.

D.D.McDonald. 1992. In S.C.Shaprio, editor, *Encyclopedia of Artificail Intelligence*, pages 983–997. Wiley, New York.

Joonhwan Lee Dongwhan Kim. 2015. : Ł . *Korean Journal of Journalism and Communication Studies*, 59:64–95.

Konstantin Nicholas Dorr. 2015. Mapping the field of algorithmic journalism. *Digital Journalism*, 4:700–722, November.

2012. Geni. http://kowey.github.io/GenI.

Bill James. 1985. In *The Bill James Historical Baseball Abstract*, pages 30–35. Free Press.

2016. Kbo official rule 2016. http://www.koreabaseball.com/file/ebook/pdf/2016rule.pdf.

2012. Kpml. http://www.purl.org/net/kpml.

2012. Naturalowl. http://www.purl.org/net/kpml.

2012. openccg. http://openccg.sourceforge.net/.

LA Times. 2014. Earthquake aftershock: 2.7 quake strikes near westwood.